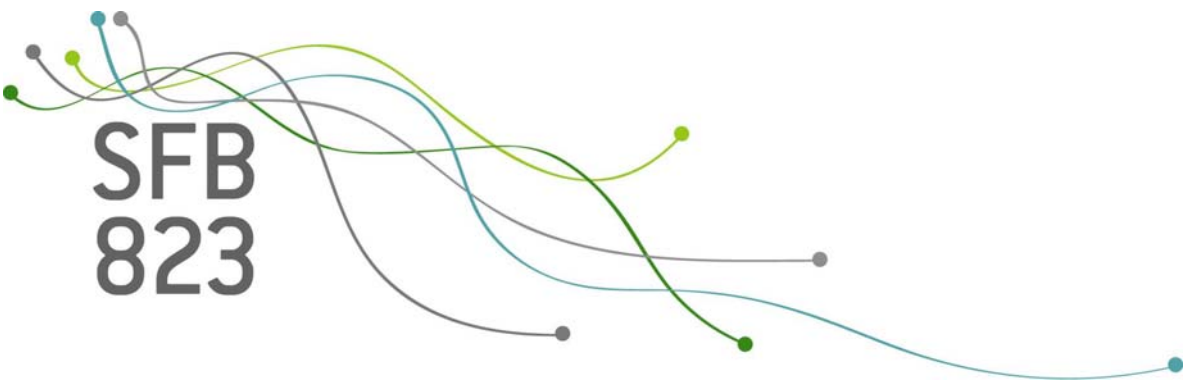
SFB 823

Comparing timbre estimation using auditory models with and without hearing loss

Klaus Friedrichs, Claus Weihs

Nr. 51/2012

Discussion Paper



COMPARING TIMBRE ESTIMATION USING AUDITORY MODELS WITH AND WITHOUT HEARING LOSS

KLAUS FRIEDRICHS AND CLAUS WEIHS

Abstract. We propose a concept for evaluating signal transformations for music signals with respect to an individual hearing deficit by using an auditory model. This deficit is simulated in the model by changing specific model parameters. Our idea is extracting the musical attributes rhythm, pitch, loudness and timbre and comparing the modified model output to the original one. While rhythm, pitch, and loudness estimation are studied in previous works the focus in this paper concentrates on timbre estimation. Results are shown for the original auditory model and three models, each simulating a specific hearing loss.

1. Introduction

For fitting and tuning a hearing aid for an individual patient as well as for fundamental research of hearing aid algorithms a method for automatic assessment based on a specific hearing deficit is very valuable. To take the knowledge of a specific hearing deficit into account Meddis proposed implementing this deficit in a widely recognized computer model of the human auditory periphery (Meddis et al., 2009). In this paper we pursue this suggestion for evaluating arbitrary complex hearing aid algorithms for music signals. This approach is shown in Figure 1. Music signals are processed by an auditory model without hearing loss and simultaneously by another model in which the simulated hearing deficit is implemented. Since the auditory model output can not be interpreted directly the musical information has to be extracted by an auralization procedure. Hence the recognized musical attributes of the two models can be compared and an evaluation of signal transformations is feasible.

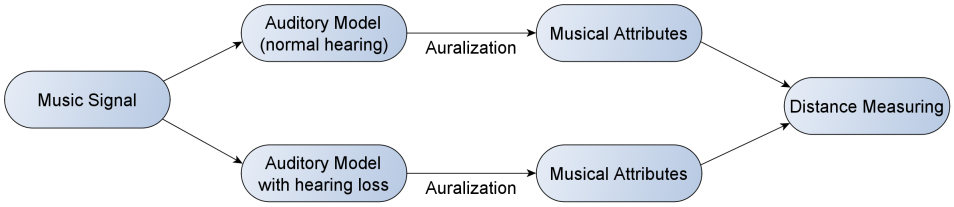


FIGURE 1. Assessment of Hearing Aid Algorithms for Music Signals

Here, auralization means a technique of decoding the auditory model output into a representation which can be understood and evaluated by humans. This can be an audio signal or in case of music the musical attributes of the signal: rhythm, pitch, loudness and timbre. For simplified auditory models the auralization task can be solved by using a genuine model inversion (Feldbauer et al. 2005). Instead, in this study an auralization approach for the model described in Meddis (2006) is presented. This model enables a more realistic implementation of different types of hearing impairments (Meddis et al., 2009). Since for Meddis' model an analytical inversion is not possible we have developed a statistical auralization approach for music signals using classification and regression methods to estimate the musical attributes: rhythm (onsets), pitch (key tones), loudness and timbre.

Meddis' auditory model is a computational simulation model of the human auditory periphery. 40 auditory nerve fibres are simulated by a 40 channels filter bank. Each channel has an individual best frequency which defines which frequencies are stimulated the most. The best frequency is between 250 Hz for the first and 7500 Hz for the 40th channel. In Figure 2 an exemplary output of the model can be seen. While the 40 channels are located on the vertical axis and the time response on the horizontal axis, the color indicates the spiking activity per second. For the auditory model with hearing loss we consider the three examples, called hearing dummies, which are described in Meddis et al. (2009). These are modified models based on the auditory model described above. The first hearing dummy simulates a bilateral moderate-severe sensory-neural hearing loss with normal middle ear function. In the model this is implemented by retaining the channel with the best frequency of 250 Hz only and by disabling the nonlinear path. The second hearing dummy simulates a bilateral, moderate, sensory-neural, sloping hearing loss with normal middle ear function.

For the model this means that the endocochlear potential is reduced from -0.1 V to -0.09 V and the channels with best frequencies above 2700 Hz are disabled. The third hearing dummy is a moderate bilateral, sensory-neural, ski-slope loss with normal middle ear function but no detectable acoustic reflex. This is implemented by disabling all channels with best frequencies above 1800 Hz.

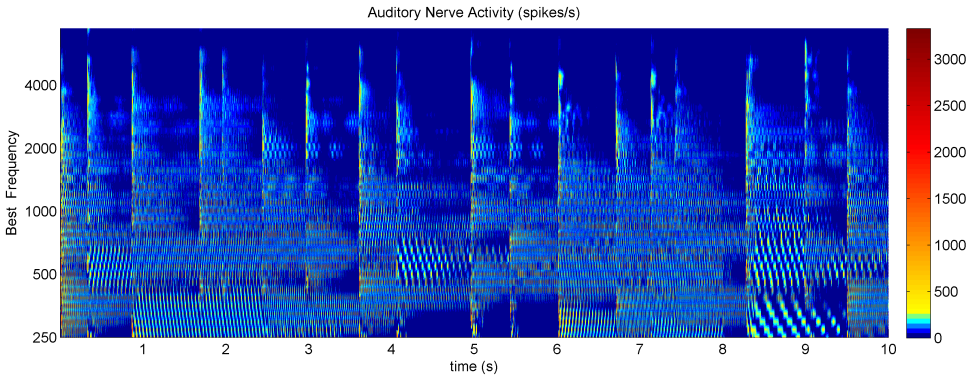


FIGURE 2. Exemplary Output of Meddis Auditory Model (Normal Hearing)

2. Auralization approach

First the auditory model output has to be separated into the single tones by using tone onset detection. Subsequently for each tone the key tone frequency, the loudness and the timbre are estimated using statistical learning procedures. During training mode the first step is skipped and instead all onset times are known.

While onset detection and frequency and loudness estimation are dealt in our previous works the focus of this paper is timbre estimation. In Bauer et al. (2012) a tone onset detection using Meddis' auditory model is proposed and compared to another approach, which uses the original signal instead. It is shown that both used representations perform altogether roughly equal. In Weihs et al. (2012) an approach for frequency detection based on Meddis' model using classification methods is introduced which solves the problem almost error free. Additionally, a method for loudness estimation for each partial tone using regression methods is presented. While in that work each partial tone is estimated separately, for the improved approach described in this paper we just need the key tone frequency and the loudness of the complete

tone, since we want to estimate also timbre. This makes these other tasks even somewhat easier.

3. Timbre estimation

Timbre is a combination of all acoustical attributes by which two musical tones which are identical in pitch, loudness and length can be distinguished (Emiroglu et al., 2007). While there are many approaches for mathematical representations for timbre many of them are highly controversial. Thus, an objective definition of timbre appears to be problematic. Therefore, we simplify timbre definition: Here, timbre distiguishes the tones of different instruments identical in pitch, loudness and length. By this we can define timbre estimation as a classification task, which can also be compared with human auditory perception.

3.1. **Experimental Design.** We use randomly generated tone sequences with known onset times which contain tones of two different musical instruments from RWC data base (Goto et al., 2003). The classification task is to identify for each tone which instrument is playing. We considered three classification tasks with different level of difficulty: piano versus clarinet, clarinet versus trumpet and piano versus guitar. While piano and clarinet can be distinguished relatively easy, the classification task for piano versus guitar is even for humans a very hard challenge, at least for the tones used in this study. For each classification task 10 randomly generated tone sequences are used. Each tone sequence consists of 20 tones, thus altogether we have 200 observations per experiment. All tones have the same duration of 0.5 seconds while the sound itensities and the pitches are randomly chosen using uniform distributions. The sound intensities have a range of [70,90] in MIDI-coding. The pitch range is dependent on the common pitch range of the respective instruments. This means [466, 4187] Hz for piano vs. clarinet, [523, 3730] Hz for clarinet vs. trumpet and [261, 2093] Hz for piano vs. guitar. For each tone one of the two instruments is randomly chosen.

All tone sequences are separately processed through Meddis' auditory model respectively each of the three hearing dummies described in Section 1. Since past observations have a significant impact on the auditory model output the same tones at different locations often produce significant different model outputs. Figure 3 shows an exemplary

output of a tone sequence piano vs. clarinet using the auditory model without hearing loss.

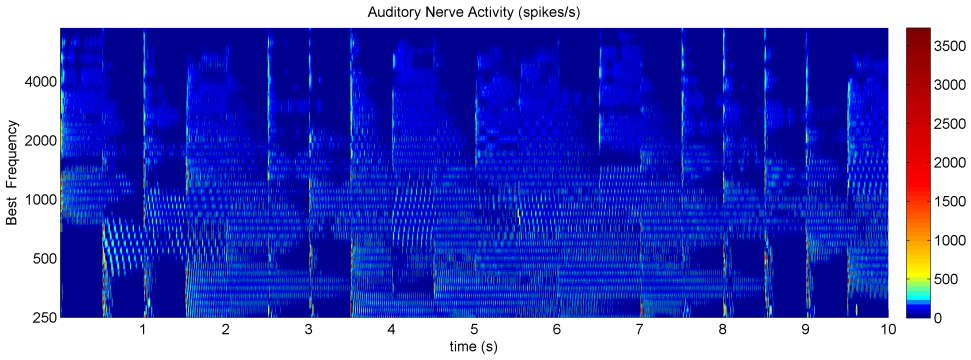


FIGURE 3. Exemplary Auditory Output of Piano vs. Clarinet (Normal Hearing)

3.2. **Feature Generation.** Contrary to other instrument recognition experiments which are based on the acoustic signal, here an additional issue is combining the outputs of up to 40 channels. In this study a relatively simple approach is chosen by using only two channel combining features. All other features are generated for each channel independently. For further simplification features are computed over the whole tone. Possibly, better results can be achieved by windowing, thereby, e.g., distinguishing the attack time from the rest. The two combining features are the average firing rate over all channels and the variance of the mean firing rates of the channels. While the first one should be more connected to loudness, the second one should be more related to pitch. Both, however, might also include some information about timbre, e.g. the amount of high frequency.

From each channel output the following features are generated. They all originate from music information retrieval for analyzing acoustic signals with respect to timbre (Lartillot et al., 2007). Thus, we can not say for sure that all of them are also an indicator for predicting timbre on the auditory model output.

a) **Shannon Entropy**

The entropy H(X) is defined as:

$$H(X) = -\sum_{i=1}^{N} p(x_i) log_2 p(x_i),$$

where X is the Discrete Fourier Transform of a signal and $p(x_i)$ is the energy share of the i-th frequency component. Shannon's

entropy is a measure of randomness in an acoustic signal and often proposed as a measure for music complexity (Madsen et al., 2006). By calculating the entropy of each channel output it can be estimated if the firing activity is just random or if it is stimulated by certain frequencies.

b) **Zero-cross (Here: "One-cross")**

Zero-cross is an indicator for noisiness of an acoustic signal by counting how often it changes the sign. Since the channels' firing activities are of cause always positive, we count how often each channel output crosses the 1-value, which might also indicate the noisiness to some degree.

c) **Roll-off 85**

Roll-off 85 is defined as the minimal value R such that

$$\sum_{i=1}^{R} x_i \geq 0.85 \sum_{i=1}^{N} x_i,$$

where x_i is the amplitude of the i-th frequency component. It estimates the amount of high frequency in an acoustic signal by calculating the frequency below which 85% of the total energy is contained. Since each channel can be considered simply as a bandpass filter, this feature might be very noisy for the auditory model output.

d) **Brightness**

This is another feature estimating the amount of high frequency in a signal. It calculates the amount of energy above 1500 Hz with respect to the total energy:

$$\sum_{i=f_{1500}}^{N} x_i / \sum_{i=1}^{N} x_i,$$

where $x_{f_{1500}}$ is the 1500 Hz frequency component. This measure should be discussed in the same way als Roll-off 85.

e) **Irregularity**

Irregularity measures the degree of variation of adjoining partials in a tone:

$$\sum_{k=1}^{N} (a_k - a_{k+1})^2 / \sum_{k=1}^{N} a_k^2,$$

where a_k is the amplitude of the k-th partial and a_{N+1} is supposed to be zero. Since the ratio of adjoining partial tones can also be seen very clearly in the auditory model output (Weihs et al., 2012), Irregularity should also indicate timbre in the auditory model output.

f) **Mel-Frequency Cepstral Coefficients (MFCCs)**

MFCCs describe the spectral shape of the sound (Davis et al., 1980). For this study the first 13 cepstral coefficients are taken as features from each channel output. However, like Roll-off 85 and Brightness, MFCCs might be very noisy due to the bandpass characteristic of the channels.

Altogether this means that we consider $2 + 40 * 18 = 722$ features in case of the original auditory model (without hearing loss).

3.3. **Results.** Using the features described above classification methods can be applied. In this study this is done by a Linear Support Vector Machine (SVM). Ten times repeated ten-fold cross validation is performed for getting significant results. These results are listed in Table 1.

TABLE 1. Misclassification rates (10 times repeated 10-fold cross validation):

	Piano vs. Clarinet	Clarinet vs. Trumpet	Piano vs. Guitar
Normal Hearing	0.0%	0.7%	4.2%
Hearing Dummy 1	31.4%	32.0%	31.5%
Hearing Dummy 2	0.6%	2.4%	6.0%
Hearing Dummy 3	2.0%	6.2%	13.4%

As expected the smallest classification error appears for the model without hearing loss and the worst one for the first hearing dummy which simulates a very strong hearing loss. Since the second hearing dummy scores better than the third one it can be supposed that a reduced endocochlear potential does not have a high impact on timbre estimation. Furthermore, the difficulties of the level of the three classification tasks is consistent with informal listening tests for all models except for the first hearing dummy, for which all tasks remarkably have nearly the same bad results.

4. Conclusion

In this paper an approach for timbre estimation respectively music instrument recognition using an auditory model is proposed. The utilized features, which are almost all originally developed for analyzing acoustic signals instead of auditory model outputs, seem already to be sufficient to produce satisfactory results. Nevertheless, we suppose that results could even be better with improved features which give more regard to the auditory model output. Additionally, in future studies we

will optimize the classification by conducting feature selection and testing different classification methods. While subsequently, a comparison with state of the art instrument recognition systems could be very interesting, our primary intention is to generate a method for automatic assessment for hearing aid algorithms. Therefore, the classifier does not need to be perfect as long as the results are consistent with the human auditory perception. However, this has to be proven by a listening test which measures the degree of correlation between statistical learning and human perception. Finally, our purpose is combining the timbre estimation with our previous studies about onset detection and pitch and loudness estimation to an overall measure, thereby gaining a method for evaluating hearing aid algorithms for music signals.

ACKNOWLEDGEMENT

This work was supported by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) of the German Research Foundation (DFG).

References

N. Bauer, K. Friedrichs, D. Kirchhoff, J. Schiffner and C. Weihs (2012): "Tone onset detection using an auditory model", SFB 823 Discussion Paper 50/12, TU Dortmund.

S.B. Davis und P. Mermelstein (1980): "Comparison of Parametric Representations for Mono-syllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech, and Signal Processing 28/4, 357-366.

S. Emiroglu, B. Kollmeier (2007): "Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions", Brain Res.

C. Feldbauer, G. Kubin and W.B. Kleijn (2005): "Anthropomorphic Coding of Speech and Audio: A Model Inversion Approach", in EURASIP Journal on Applied Signal Processing, Volume 2005.

M. Goto, H. Hashiguch, T. Nishimura, and R. Oka (2003): "RWC Music Database: Music Genre Database and Musical Instrument Sound Database", Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003), 229-230.

O. Lartillot, P. Toiviainen (2007): "MIR in Matlab (II): A toolbox for musical feature extraction from audio", in International Conference on Music Information Retrieval.

S. T. Madsen, G. and Widmer (2006): "Music complexity measures predicting the listening experience", in Proceedings of the 9th International Conference on Music Perception and Cognition, Bologna.

R. Meddis (2006): "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model", Journal of the Acoustical Society of America 119, 406-417.

R. Meddis, W. Lecluyse, C.W. Tan, and M.R. Panda (2009): "Beyond the audiogram: identifying and modelling patterns of hearing deficits", in The Neurophysiological Bases of Auditory Perception, Proc. of the International Symposium on Hearing.

C. Weihs, K. Friedrichs und B. Bischl (2012): "Statistics for hearing aids: Auralization", in J. Pociecha, R. Decker (Hrsg.): Data Analysis Methods and its Applications, 183-196.

Chair of Computational Statistics, TU Dortmund
E-mail address: {friedrichs, weihs} @statistik.tu-dortmund.de